AUTOMATIC DOCUMENT CLASSIFICATION USING TEXT AND IMAGES

FIELD OF THE INVENTION

The present invention relates to document management. More particularly, the present invention relates to automatic classification of documents using both text and images.

BACKGROUND OF THE INVENTION

Typically, electronic documents are stored in a hierarchical structure of directories/folders during or after creation. For example, when a user creates a document using a word processing application, the user saves the document to a directory or sub-directory of a storage device, such as a hard drive. Similarly, electronic documents that are generated from physical documents, for example, by scanning the physical document are stored by the user generating the new electronic document.

As electronic documents become more prevalent in offices and homes, conversion of documents from physical form to electronic form may become more common. Individual assignment of directories for each document can be time consuming and tedious. Also, as physical documents are manipulated using copiers, facsimile machines and printers, electronic copies may be saved for later retrieval. Individual electronic storage of each physical document that is manipulated can quickly become time consuming.

What is needed is a method and apparatus for automatic classification and storage of electronic documents.

SUMMARY OF THE INVENTION

A method and apparatus for automatic document classification using text and images is described. An original directory structure is analyzed to determine classification characteristics of documents previously stored. In one embodiment, a shadow, or mirror, directory structure is generated based on the original directory structure. Documents are automatically classified and stored based on the characteristics of the original directory structure.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings in which like reference numerals refer to similar elements.

Figure 1 is one embodiment of a document processing device.

Figure 2 is one embodiment of a document processing system having multiple document processing devices.

Figure 3 is one embodiment of a flow diagram of a process for generating a mirror directory structure.

Figure 4 is one embodiment of a flow diagram of a process for classifying electronic documents.

Figure 5 is one embodiment of a flow diagram of a process for performing textual analysis on an electronic document.

Figure 6 is one embodiment of a flow diagram of a process for performing graphical analysis on an electronic document.

Figure 7A-7D illustrate exemplary web pages to traverse a mirror hierarchy.

DETAILED DESCRIPTION

A method and apparatus for automatic document classification using both text and images is described. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention can be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid obscuring the present invention.

Some portions of the detailed descriptions which follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like.

It should be borne in mind, however, that all of these and similar

terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

The present invention also relates to apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general purpose computer selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, and magneto-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus. The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general purpose machines may be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these machines will appear from the description below. In addition, the present invention is not

described with reference to any particular programming language. It will be appreciated that a variety of programming languages may be used to implement the teachings of the invention as described herein.

Overview

Briefly, a method and apparatus for automatic document classification based on both text and image is described. A document is analyzed based on textual content as well as visual appearance to obtain both a text-based classifier and an image-based classifier, respectively. These classifiers may be used to characterize, or profile, documents.

Based on a document's textual and visual content, the document is automatically stored in the one or more directories (or folders) of a document hierarchy in which the document would most likely be stored by a user if the document were placed there manually by the user. Determination of the most likely directories is based on an analysis of previously stored documents stored by a user in those and other directories. In one embodiment, the directories are components of a mirror directory structure, which is generated by classifying documents in a pre-existing directory structure, such as the user's hard drive. Text-based and image-based classifiers are generated for the documents in the individual directories. These classifiers are combined to obtain to profile describing documents in that directory.

By storing the new document automatically, the user is relieved of the duty of manually selecting a directory for the new document. Also, because the storage location of a document is based on the user's previously created

storage arrangement, the user should be able to more easily locate the document.

System Description

Figure 1 is one embodiment of a document processing device. Document processing device 100 includes bus 101 or other communication device for communicating information, and processor 102 coupled to bus 101 for processing information. Document processing device 100 further includes random access memory (RAM) or other dynamic storage device 104 (referred to as main memory), coupled to bus 101, for storing information and instructions to be executed by processor 102. Main memory 104 also can be used for storing temporary variables or other intermediate information during execution of instructions by processor 102. Document processing device 100 also includes read only memory (ROM) and/or other static storage device 106 coupled to bus 101 for storing static information and instructions for processor 102. Data storage device 107 is coupled to bus 101 for storing information and instructions.

Data storage device 107 such as a magnetic disk or optical disc and corresponding drive can be coupled to document processing device 100. Document processing device 100 can also be coupled via bus 101 to display device 121, such as a liquid crystal display (LCD), for displaying information to a user. Input device 125 allows a user of document processing device 100 to

provide input and control. Input device 125 can be, for example, a keyboard, a keypad, a mouse, a trackball, a trackpad, a touch-sensitive screen, etc.

The present invention is related to the use of document processing device 100 to automatically classify documents using both text and images. According to one embodiment, automatic classification is performed by document processing device 100 in response to processor 102 executing sequences of instructions contained in memory 104. Execution of the sequences of instructions contained in memory 104 causes processor 102 to automatically classify documents based on text and images, as will be described hereafter. Instructions are provided to main memory 104 from a storage device, such as magnetic disk, CD-ROM, DVD, via a remote connection (e.g., over a network), etc. In alternative embodiments, hard-wired circuitry can be used in place of or in combination with software instructions to implement the present invention. Thus, the present invention is not limited to any specific combination of hardware circuitry and software.

Document processing device 100 can be a computer system in which documents are generated with an application program such as a word processing program, electronic mail program, spreadsheet program etc. Document processing device 100 can also be a copier, facsimile (fax) machine, or printer that stores copies of documents processed. For example, a copier can store images of documents copied. A fax machine can store images of

documents sent or received. A printer can store copies of the documents printed.

In one embodiment, document processing device 100 is the ImageHunter™ imaging system available from Ricoh Company, Ltd. of Tokyo, Japan. In such an embodiment, document processing device 100 is an image-based filing system that digitally stores paper documents into electronic format. Conversion of paper documents into filed images, increases accessibility of information compared to manual-based systems such as cabinets, storage boxes, etc.

Figure 2 is one embodiment of a document processing system having multiple document processing devices. The system of Figure 2 is described in terms of multiple document processing devices interconnected by a network. However, a single such device can provide a document processing system.

In one embodiment, network 200 is a local area network that interconnects multiple document processing and other computing devices. However, other types of networks can be used. For example, network 200 can be the Internet or other wide area network.

Copier 210 is a document copying device that can store documents in database 240 or other storage device, either internal or external to copier 210. Copier 210 is coupled to network 200 to communicate images of documents copied as well as control and other information. Documents can be communicated to other devices coupled to network 200 for further processing or other purposes. In one embodiment, documents copied by copier 210 are

stored in database 240 for later retrieval. By storing documents that have been processed the original paper copy of the document is no longer necessary. If the document is needed an electronic version of the document can be retrieved from database 240.

Fax machine 220 is also coupled to network 200. Fax machine 220 stores copies of documents sent and received in database 240 or other storage device, which can be internal or external to fax machine 220. Documents can, for example, be retrieved from database 240 directly and sent via fax machine 220 without the need of a physical document. Similarly, printer 250 can print documents created by devices coupled to network 200 or documents retrieved from database 240 or other storage device.

Computer system 230 can be any type of computer system. In one embodiment, a hard disk (not shown in Figure 2) of computer system 230 is used to determine the organization of new electronic documents. Because the hard disk of a computer system is organized in a manner that is logical to a user of computer system 230, storage of new electronic documents in the same or similar organization allows new documents to be placed automatically and to be easily retrievable by the user. Automatic organization and storage of new electronic documents is described in greater detail below. Alternatively, a directory structure of database 240 or other storage device can be used to organize documents that are automatically stored.

Printer 250 can be coupled to network 200. Printer 250 can be, for example, a printer that stores data in database 240. Printer 250 provides

physical copies of electronic documents and can also store electronic documents on database 240. In one embodiment, a single device provides the functionality of copier 210, fax machine 220 and printer 250.

In order to facilitate the classification scheme for documents as described herein, a document hierarchy is first created which provides a partitioning of the document space based on a user's organization of that space. In one embodiment, this partitioning is accomplished by creating a directory structure that mirrors the directory-based storage memory already in use by an individual. In one embodiment, a system extracts an organization that a user has already applied to files on a computer system.

Figure 3 is one embodiment of a flow diagram of a process for generating a mirror directory structure. As described below, in one embodiment, the present invention analyzes a pre-existing memory hierarchy, such as, for example, a directory structure on user's hard drive. Other pre-existing directory structures include Web page bookmarks, document storage directories, etc. and these may also be used to generate the mirror directory structure. The processing of Figure 3 is performed by processing logic. The processing logic may comprise software running on general purpose or dedicated computer system or machine, or may comprise dedicated hardware, or a combination of both.

Referring to Figure 3, processing logic initially copies the pre-existing directory structure (processing logic 310). In one embodiment, each directory having anything stored therein such as, for example, files, folders,

applications, subdirectories, etc. is copied into a duplicate directory structure having the same directory structure as the original. In one embodiment, the new directory structure may be referred to as the "mirror directory structure" and is a copy of the original directory structure at the time the mirror directory structure is created. In an alternate embodiment, only those portions of the directory with a minimum number of files, folders, applications, etc. are copied. The minimum number may be any number set by user or designer choice, such as, for example, 3, 4, or 5 documents, etc.

In one embodiment, all of the directories (e.g., folders) are located by recursively descending the file system hierarchy. Processing logic creates a list of all of these directories, selects each directory, and examines the files contained in the directory. Processing logic filters files in each of these directories for their content. Files that comprise text, including aschii text, postscript, pdf, etc., are labeled as having text features, while those containing an image, such as those that can be rendered (e.g., postcript, tiff, etc.) are also labeled as having image features. The labeling may be later used to further classify the individual documents, with the directory name labeling the class of documents store therein. The labeling may also be used as an index into a database storing all of the documents. A threshold may be set where a minimum number of document based files must be found before a directory is maintained for a type of document. In this manner, the total number of directories is reduced to a set that contains the documents on the hard drive.

In one embodiment, the process is performed controllably through one or more scripts that filter through the user's hard drive. The process may be capability controlled by the user and is analogous to running a virus checker on the hard drive. The script is run initially when the user starts to build their own hierarchical document database.

Once a mirror hierarchy is generated, the system is trained to recognize the types of documents in each selected directory. Processing logic analyze documents found within the original directory structure (processing block 320). In one embodiment, analysis includes both textual analysis and graphical analysis of documents contained therein. Various embodiments of textual analysis and graphical analysis are described in greater detail below. Textual and graphical analysis are used to characterize the subdirectories within the original directory.

Processing logic then builds a directory profile using the results of the analysis (processing block 330). In one embodiment, a directory profile is built for each sub-directory of the mirror directory structure. The directory profile is used to match new documents with their appropriate storage locations within the mirror directory structure based on the classification of documents already stored in that portion of the directory. Thus, the results of performing textual and graphical analysis on documents in the original directly is to obtain a classifier or classification for each subdirectory. Later, when determining where to store a particular document, the same textual and

graphical analysis is applied to that document and the results are compared to the existing classifiers to determine where to store the document.

In one embodiment, the mirror directory structure can be periodically updated based on modifications to the original directory. If a user moves documents between directories in the original directory, the classifiers for directories in the mirror directory may be updated to better reflect the user's new idea of classification. Note that this may be done on a periodic basis or only when an actual change has occurred. Also, all the classifiers may be updated or only those associated with the document(s) moved by the user. The user may perform the update by re-running a script. Therefore, as a user changes the hard drive directory structure, a new structure for the user's documents may be learned by the system.

Figure 4 is one embodiment of a flow diagram of a process for classifying electronic documents. The classifying of the electronic document allows the document to be stored in a database or other memory transparently to the user. Processing logic, as described above, performs the process. In one embodiment, a script takes the document and runs an auto-filing process where the newly arrived document gets placed in the mirror hierarchy transparently to the user.

Referring to Figure 4, processing logic receives an electronic document (processing block 410). The electronic document can be received, for example, from a scanner, a copier, a fax machine, a computer system, etc. The electronic document can be a pre-existing electronic document that is to be

automatically classified and stored, or the electronic document can be a document that has been recently converted from physical to electronic form, for example, with a scanner. The text of any scanned image may undergo optical character recognition, and both the image data and the text obtained via optical character recognition are saved.

Processing logic performs textual analysis on the electronic document (processing block 420). Textual analysis is described in greater detail below, for example, with respect to Figure 5. Processing logic uses the results of the textual analysis to build a textual document profile (processing block 430). The textual document profile is a text-based classifier.

In parallel, processing logic performs graphical analysis on the electronic document (processing block 440). Graphical analysis is described in greater detail below, for example, with respect to Figure 6. Processing logic uses the results of the graphical analysis to build the graphical document profile (processing block 450). The graphical document profile is an image-based classifier.

In the example of Figure 4, the textual analysis and graphical analysis are described as being performed in parallel; however, the textual analysis and graphical analysis can be performed in a sequential manner.

Processing logic uses the textual document profile and graphical document profile to build a document profile (processing block 460). In one embodiment, a Borda Count method is used to combine the results of the textual analysis (the text-based classifier) and the graphical analysis (the

image-based classifier) to build the document profile. For more information on the Borden Count method, see Tim Ho, et al., "Decision Combination in Multiple Classifier Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 1, January 1994.

The Borda Count method assigns points in a descending manner to each choice and then sums the points to determine one or more preferred choices. For example, the directory having the best match based on textual analysis can be assigned three points, the directory having the second best match can be assigned two points, and the directory having the third best match can be assigned one point. Similar assignments are made based on graphical analysis. Of course, points can be assigned in a different manner (e.g., the top five matches, the top seven matches).

The points assigned based on the textual and graphical analysis are combined to determine overall matches. In one embodiment, each electronic document is stored in three directories based on the results of the textual and graphical analysis. Thus, the matches having the three highest totals based on the Borda Count indicate the directories in which the electronic documents are stored. Other techniques for combining the results of the textual analysis and graphical analysis can be used to build the document profile based on both textual analysis and graphical analysis. For example, combining may be done by logistic repression or the highest rank method such as described in Tim Ho, et al., "Decision Combination in Multiple

Classifier Systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 1, January 1994.

The document profile is used to determine where, within the mirror directory structure, the electronic document should be stored. In one embodiment, multiple storage locations are selected based on document profile. Processing logic then stores the document in the one or more locations determined in response to the document profile (processing block 470).

Figure 5 is one embodiment of a flow diagram of a process for performing textual analysis on an electronic document. The textual analysis can be used both for analysis of the documents in the original directory structure as well as electronic documents to be automatically classified and stored. Processing logic, as described above, performs the process of Figure 5.

Referring to Figure 5, processing logic analyzes the text of the electronic document (processing block 510). In one embodiment, this analysis includes extracting the text and processing the text for characteristic words. In one embodiment, stop words (e.g., "a", "the", "but") are removed from extracted text and the remaining words are stemmed. A table is built based on the frequencies of the remaining words. These words are the least frequently used words and the occurrence rate associated with each search word is used as a basis of the document profile.

A second characteristic that can be used to classify electronic documents is the ratio of the number of words in a document to the number of lines in

the document. Of course, other types of textual analysis can also be used. Other textual characteristics that may be used in textual analysis include the transition probability of word lengths, such as, for example, the textual analysis described in U.S. Patent ~~Application Serial No. _____, entitled~~ 5,909,680, entitled "Document Categorization by Word Length Distribution Analysis," ~~filed~~ issued June 1, 1999, ~~September 9, 1996~~ and character N-gram probabilities such as described in U.S. Patent 5,418,951, entitled "Method of Retrieving Documents that Concern the Same Topic, issued May 23, 1995.

Processing logic builds a document profile based on characteristics determined through the textual analysis (processing block 520). In one embodiment, after the appropriate textual characteristics have been determined, a naive Bayes classifier is used to match the one or more textual characteristics to pre-existing textual characteristics corresponding to directories. One embodiment of a Bayes classifier is described in greater detail in "Machine Learning," McGraw-Hill Companies, Inc., First Edition, 1997. Alternatively, a more sophisticated classifier, such as a neural network, can be used.

Figure 6 is one embodiment of a flow diagram of a process for performing graphical analysis on electronic document. The graphical analysis comprises processing the image data for features. As with textual analysis described above, graphical analysis can be used both for analysis of documents in the original directory structure and electronic documents to be

automatically classified and stored. Processing logic, as described above, performs the processing shown in Figure 6.

Processing logic analyzes a graphical representation of the new electronic document (processing block 610). In one embodiment, the analysis includes generating a copy of the electronic document that represents the document as a known graphical format (e.g., PostScript, PDF). Processing logic extracts features. Those features may be based on texture, statistical moments and for distribution of connected components in the document. In alternative embodiments, features may also include one or more of corner points, edges, and line segments. Processing logic then represents the electronic document as a point set (processing block 620), in which a point identifies each intersection of two or more lines. Other graphical representations can also be used. For example, lines as described in a PostScript file can be used.

Processing logic builds a document profile based on the graphical representation of the new document (processing block 630). In one embodiment, the density of points in the point set corresponding to pre-defined areas of the electronic document is used to build the document profile. Other graphical document profiles can also be built using additional or different graphical representations. The document profile is compared with one or more pre-existing directory profiles to determine one or more locations for the electronic document to be stored.

Using a nearest neighbor classifier, documents that are closest to the document being added are found and a set of candidate labels are assigned. Another approach that may be used is to match the document against the centroids of the previously determined clusters. A "cluster" refers to have a group of documents. Described herein is a top-down method of building clusters in which one document at a time is added to an existing hierarchy.

In a bottom-up cluster building approach, the system is given a set of previously unclassified documents. Feature vectors (containing either image-based or text-based features) are extracted from a document. A standard hierarchical clustering algorithm, such as described in A.K. Jain & R.C. Dubes, "Algorithms for Clustering Data," Prentice Hall, 1988, is applied to the feature vectors. This organizes the documents in a hierarchical tree structure. Terminal nodes in the tree correspond to groups of "similar" documents.

A single centroid vector can be generated for such a cluster in the standard way. Image-based feature vectors are typically of fixed length. A centroid feature vector is generated by adding individual elements from the members of a cluster and dividing by cluster size. Text-based feature vectors, which contain words and their frequencies, are typically of different lengths. The centroid vector is generated in a similar way. It contains a number of elements equal to the total number of unique words in the documents in the cluster.

An unknown document can be matched to such a cluster hierarchy by computing the Euclidean distance between its feature vector and the centroid

feature vectors for the terminal nodes in the hierarchy. The unknown document can be added and assigned to the N nodes that minimize the Euclidean distance.

Note that these techniques apply to text as well as graphics.

User interaction may be allowed with the documents in the mirror hierarchy. The user interaction may occur when a document needs to be retrieved. In retrieving a document, the user need only rely on remembering the organization of their hard drive since documents are added to the mirror hierarchy based on the training off the users hard drive.

A web based viewer may be used to traverse the mirror hierarchy. Figures 7A-7D illustrate an example set of web pages to navigate from the top level of a hierarchy to a "leaf" node that contains a group of similar documents. These contain both similar images and documents related by their textual content.

In the foregoing specification, the present invention has been described with reference to specific embodiments thereof. It will, however, be evident that various modifications and changes can be made thereto without departing from the broader spirit and scope of the invention. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.